

Optimizing the quantity/quality trade-off in connectome inference*

Carey E. Priebe¹, Joshua Vogelstein¹, Davi Bock²

¹JHU AMS & ²HHMI Janelia

October 13, 2011

Abstract

We demonstrate a meaningful prospective power analysis for an (admittedly idealized) illustrative connectome inference task. Modeling neurons as vertices and synapses as edges in a simple random graph model, we optimize the trade-off between the number of (putative) edges identified and the accuracy of the edge identification procedure. We conclude that explicit analysis of the quantity/quality trade-off is imperative for optimal neuroscientific experimental design. In particular, identifying edges faster/more cheaply, but with more error, can yield superior inferential performance.

Introduction

Statistical inference on graphs begins with modeling graph-valued observations $G = (V, E)$, where $V = \{1, \dots, n\}$ is the vertex set and $E \subset V \times V$ is the edge set (connections between vertices), via a random graph model $\mathbb{G} \sim P_{\theta_0} \in \mathcal{P} = \{P_{\theta} : \theta \in \Theta\}$. The parameter θ governs the distribution P_{θ} over the collection \mathcal{G}_n of possible graphs on n vertices, and the parameter set Θ indexes the distributions in our model. Inference then proceeds via estimation or hypothesis testing regarding the true but unknown parameter value $\theta_0 \in \Theta$.

Statistical inference on connectomes – graphs representing brain structure – involves positing a probabilistic model for the connectome and deriving desirable properties of a statistic (a function of G or \mathbb{G}) with respect to neuroscientific questions regarding θ_0 .

For example, electron microscopy (EM) and magnetic resonance (MR) imaging technology can produce high-resolution connectome data. In EM, the connectome is the graph obtained by representing neurons as vertices and synapses as edges. In MR, the vertices represent voxels or neuroanatomical regions and the edges represent functional, effective, or structural connectivities. It is estimated that a human or primate cortical column contains approximately 100,000 neurons, each with about 10,000 connections, yielding approximately one billion synapses. Thus one could hope to observe a massive graph and perform inference thereon, yielding fundamentally important neuroscientific understanding.

*This work is partially supported by the Research Program in Applied Neuroscience.

However, given the imaging data, one must estimate the graph. Typically, these brain-graphs are obtained by tedious and time consuming manual annotation. In Bock et al. [2011], approximately nine expert-human months were required to find 250 synapses in EM imagery of the mouse primary visual cortex – about 1 synapse per expert-human day. At that rate, it would take nearly 300 million expert-human years to recover the full induced subgraph of a primate cortical column. It is possible that annotating more quickly – and more errorfully – would yield superior statistical inference. Indeed, regardless of the scale of the connectome, or the imaging technology, there is an inherent quantity/quality trade-off in statistical connectomics.

In this manuscript, we present an (admittedly idealized) illustrative setting in which we optimize the quantity/quality trade-off analytically, demonstrating that identifying brain-graph edges faster/more cheaply, but with more error, can yield superior inference in statistical connectomics. We describe a very simple brain-graph model and a correspondingly simple error model to explicate how the quantity/quality trade-off impacts the power of a particular hypothesis test.

Connectomic Motivation

The connections made by cortical brain cells are anatomically nanoscopic, yet each cell in the cortex has several centimeters of local anatomical “wiring” (Braitenberg and Schüz [1998]). This wiring packs the cortical volume essentially completely. Bock et al. [2011] recently characterized the *in vivo* responses of a group of cells in mouse visual cortex, then imaged a volume of brain containing the cells using a custom-built high throughput electron microscopy (EM) camera array. Each voxel in the resulting data set occupies about $4 \times 4 \times 45$ cubic nanometers of brain; the 10 teravoxel volume spans $450 \times 350 \times 50$ cubic micrometers. The imaged volume is of sufficient size and resolution that they were able to trace the local connectivity of the physiologically characterized cells. One can therefore record what cells in the brain are doing and then trace their connectivity – a combination which could enable a new level of understanding of cortical circuits to be achieved.

Model & Hypotheses

Let \mathbb{G} be an independent edge stochastic block model on n vertices.

Vertices represent neurons. We denote by \mathcal{E} the collection of n_E excitatory neurons and by \mathcal{I} the collection of n_I inhibitory neurons. Thus the vertex set V is decomposed as the disjoint union $V = \mathcal{E} \cup \mathcal{I}$. Let $n = |V| = |\mathcal{E}| + |\mathcal{I}| = n_E + n_I$ with $n_E = \lambda n$ and $n_I = (1 - \lambda)n$ for some $\lambda \in (0, 1)$.

Edges represent synapses. We consider loopy graphs – a neuron may connect to itself. We consider the undirected edge case for simplicity; the directed and multigraph cases follow *mutatis mutandis*.

The block model structure is such that

$$\begin{aligned} P[u \sim v] &= p_{EE} \quad \text{for } u, v \in \mathcal{E}, \\ P[u \sim v] &= p_{II} \quad \text{for } u, v \in \mathcal{I}, \\ P[u \sim v] &= p_{EI} = p_{IE} \quad \text{otherwise.} \end{aligned}$$

That is, the probability that two excitatory neurons connect to one another is given by random graph model parameter p_{EE} , the probability that two inhibitory neurons connect to one another is given by p_{II} , and the probability that an excitatory neuron connects to an inhibitory neuron is given by $p_{EI} = p_{IE}$ (necessarily equal if and only if we are considering the undirected case).

We assume that the excitatory-excitatory connection rate p_{EE} and the inhibitory-inhibitory connection rate p_{II} are equal. We propose to test the hypothesis that this common rate $p_{EE} = p_{II}$ is equal to the excitatory-inhibitory connection rate p_{EI} :

$$\begin{aligned} H_0 : \quad & p_{EE} = p_{II} = p_{EI} \\ & vs. \\ H_A : \quad & p_{EE} = p_{II} < p_{EI} . \end{aligned}$$

The available data are an observed collection of *putative* or *errorful* edges.

Data

Ideally, we would observe the entire induced subgraph $G = \Omega(V; G^*)$ for our imaged volume, where G^* is the entire brain graph (connectome). Practically, since identifying edges is expensive, for large n we will observe a subgraph – a subset of edges. For $i = 1, \dots, z$, we define the random variable X_i representing a perfect edge observation via the “tracing algorithm” given by

- (1) a neuron: choose a vertex v_i uniformly at random from V .
- (2) a synapse: choose an edge $v_i \sim \cdot$ uniformly at random from among edges incident to v_i .
- (3) the post-synaptic neuron: identify vertex w_i for $v_i \sim w_i$.
- (4) the nature of the synapse: $X_i = I\{v_i, w_i \in \mathcal{E} \text{ or } v_i, w_i \in \mathcal{I}\}$.
 - If w_i is not in our imaged volume, or if the edge = axon-synapse-dendrite goes outside our imaged volume so that we cannot trace it, we try again (with the same i).
 - If $v_i \sim w_i$ is by chance previously identified, we try again (with the same i).

However, the algorithm above requires *perfect* tracing. Unfortunately, perfect edge observations are expensive, even for a subset of edges. Instead, with error probability $\varepsilon \in [0, 1]$ we fail to correctly identify w_i and instead identify a randomly chosen vertex \tilde{w}_i in Step (3) above, yielding

- (3') identify \tilde{w}_i for $v_i \sim \tilde{w}_i$; with probability $(1 - \varepsilon)$ $\tilde{w}_i = w_i$, otherwise \tilde{w}_i is random.
- (4') $\tilde{X}_i = I\{v_i, \tilde{w}_i \in \mathcal{E} \text{ or } v_i, \tilde{w}_i \in \mathcal{I}\}$.

Thus $\tilde{X}_i = 1$ if the i^{th} (putative) edge is either an excitatory-excitatory connection or an inhibitory-inhibitory connection, and $\tilde{X}_i = 0$ if the i^{th} (putative) edge is an excitatory-inhibitory connection. The edge tracing algorithm generates z such ε -errorful edges, yielding an errorful subgraph observation model.

Trade-off

Presumably, the expense of edge tracing is increasing in putative edge count z for fixed edge tracing error ε and decreasing in ε for fixed z . For fixed resources (imaging resolution and/or manual or automatic edge tracing resources) the trade-off of interest is z vs. ε – quantity vs. quality. The number of ε -errorful edges that will be traced, $z = h(\varepsilon)$, is an increasing function of ε . We derive below the optimal operating point of the quantity/quality trade-off in a particular connectome inference task.

Of course, committing more resources (higher-resolution imaging and/or additional edge tracing resources) should yield larger z for the same ε or smaller ε for the same z ; a prospective power cost/benefit analysis can be performed to aid in the decision regarding commitment of

resources. Still, the quantity/quality trade-off is an essential component of any such cost/benefit analysis, since one would want to consider the optimal quantity/quality operating point for each level of resource commitment.

Inference

We derive the expression for

$$\begin{aligned} P[\tilde{X}_i = 1] = p_{\tilde{X}} &= p_{\tilde{X}}(n, \lambda, p_{EE}, p_{EI}, \varepsilon) \\ &= (1 - \varepsilon) \left(\frac{\lambda p_{EE} n_E}{p_{EE} n_E + p_{EI} n_I} + \frac{(1 - \lambda) p_{II} n_I}{p_{II} n_I + p_{IE} n_E} \right) + \varepsilon(2\lambda^2 - 2\lambda + 1). \end{aligned}$$

Note that under H_0 the value of $p_{\tilde{X}}(n, \lambda, p_{EE}, p_{EI}, \varepsilon)$ is independent of the value of $p_{EE} = p_{EI}$, and that $p_{\tilde{X}}$ is *smaller* under the alternative hypothesis $p_{EE} < p_{EI}$ than under the null. Since we have (approximately) independent random variables $\tilde{X}_i \sim \text{Bernoulli}(p_{\tilde{X}})$, we reject for small values of the test statistic $\tilde{X}_z = \frac{1}{z} \sum_{i=1}^z \tilde{X}_i$ based on having observed z errorful edges. Assuming independent errors, this test is uniformly most powerful (UMP). Applying the central limit theorem under both H_0 and H_A yields a large n large z normal approximation for the power of the level α test,

$$\begin{aligned} P[\tilde{X}_z < c_\alpha | H_A] = \beta_{z,\varepsilon} &= \beta_{z,\varepsilon}(n, \lambda, p_{EE}, p_{EI}, \alpha) \\ &= \Phi \left(\frac{p_{\tilde{X}}^0 (1 - p_{\tilde{X}}^0) \Phi^{-1}(\alpha) + \sqrt{z} (p_{\tilde{X}}^0 - p_{\tilde{X}}^A)}{p_{\tilde{X}}^A (1 - p_{\tilde{X}}^A)} \right), \end{aligned}$$

where $p_{\tilde{X}}^0$ and $p_{\tilde{X}}^A$ denote the value of the Bernoulli parameter $p_{\tilde{X}}(n, \lambda, p_{EE}, p_{EI}, \varepsilon)$ given above under H_0 (whatever be the value of $p_{EE} = p_{EI}$) and under H_A (for specific values of $p_{EE} < p_{EI}$), respectively.

Perfect edge tracing ($\varepsilon = 0$) for z edges yields power $\beta_{z,0} > \alpha$. Errorful edge tracing ($\varepsilon > 0$) for z putative edges yields power $\beta_{z,\varepsilon} < \beta_{z,0}$. As expected, more error yields less power for fixed putative edge count z : $\varepsilon_1 < \varepsilon_2$ implies $\beta_{z,\varepsilon_1} > \beta_{z,\varepsilon_2}$ and $\beta_{z,1} = \alpha$ for any z . Furthermore, more edges yields more power for fixed edge tracing error rate ε : $z_1 > z_2$ implies $\beta_{z_1,\varepsilon} > \beta_{z_2,\varepsilon}$ for any ε . However, we can identify equivalent sample size z'_ε such that $\beta_{z'_\varepsilon,\varepsilon} \approx \beta_{z,0}$. Thus, if errorful edge tracing is sufficiently less expensive so that we can trace more than z'_ε errorful edges compared to just z perfect edges (which is plausible since perfect edge tracing is expensive while errorful edge tracing should be less so) then inferential performance based on an errorful subgraph will be superior to inferential performance based on a perfect subgraph. This suggests that we may benefit from optimizing the quantity/quality trade-off with respect to power for fixed resources.

Example

A mouse cortical column (the existence of which is admittedly the subject of neuroscientific debate; we proceed with an illustrative example regardless) has approximately 10,000 neurons. With parameter values $n = 10000, \lambda = 0.9, p_{EE} = p_{II} = 0.1, p_{EI} = 0.2$ for the random graph model \mathbb{G} , we expect the graph to have roughly 6 million edges total in the induced subgraph. High-accuracy manual edge tracing results in approximately one edge per expert per day. For

these parameter values, testing at level $\alpha = 0.05$ yields

$$\begin{aligned}\beta_{50,0} &\approx 0.429, \\ \beta_{50,0.5} &\approx 0.196, \\ \beta_{250,0.5} &\approx 0.488.\end{aligned}$$

Thus our prospective power analysis demonstrates that less expensive errorful edge tracing can be inferentially superior to more expensive perfect edge tracing: if we can trace $z = 50$ edges perfectly ($\varepsilon = 0$) we obtain power $\beta_{50,0} \approx 0.429$ (compared to degraded power $\beta_{50,0.5} \approx 0.196$ with the same number of (putative) edges ($z = 50$) and 50% edge tracing error ($\varepsilon = 0.5$)), while if we can trace $z = 250$ putative edges with 50% edge tracing error ($\varepsilon = 0.5$) we obtain significantly improved power $\beta_{250,0.5} \approx 0.488 > \beta_{50,0} \approx 0.429$. The equivalent sample size for this example is $z'_{0.5} = 178$, so that $\beta_{178,0.5} \approx \beta_{50,0} \approx 0.429$; thus tracing more than 178 50%-errorful putative edges yields higher power than that obtained with 50 errorless edges.

Extending this example, we assume that $z = h(\varepsilon)$. That is, the number of (errorful) putative edges that we can trace with edge tracing error ε is given by some (increasing) function h of ε . Thus the power $\beta(\varepsilon)$ obtained when using the edge tracing algorithm engineered to produce $z = h(\varepsilon)$ putative edges with edge tracing error ε is given by

$$\beta(\varepsilon) = \Phi(g(\varepsilon))$$

where

$$g(\varepsilon) = \frac{p_{\tilde{X}}^0(\varepsilon)(1 - p_{\tilde{X}}^0(\varepsilon))\Phi^{-1}(\alpha) + h(\varepsilon)^{1/2}(p_{\tilde{X}}^0(\varepsilon) - p_{\tilde{X}}^A(\varepsilon))}{p_{\tilde{X}}^A(\varepsilon)(1 - p_{\tilde{X}}^A(\varepsilon))}.$$

Assuming that h is differentiable with respect to ε on $[0, 1)$, we obtain

$$\frac{\partial \beta}{\partial \varepsilon} = \phi(g(\varepsilon))g'(\varepsilon).$$

Then we evaluate the sign of $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon_0}$ at the current edge tracing algorithm operating point ε_0 ; $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon_0} > 0$ implies less expensive more errorful (larger ε) edge tracing (resulting in larger z) will yield increased power, while $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon_0} < 0$ implies that inference will improve with more accurate but more expensive edge tracing (resulting in fewer putative edges). Finding ε^* such that $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon^*} = 0$ will (after checking appropriate side conditions) yield optimal power $\beta^* = \beta(\varepsilon^*)$. To continue with our example, we consider for illustration

$$z = h(\varepsilon) = 50 + \frac{200}{\sin(\pi/4)} \sin(\varepsilon\pi/2),$$

designed to give $h(0) = 50$, $\beta(0) \approx 0.429$ and $h(1/2) = 250$, $\beta(1/2) \approx 0.488$ for consistency with our running example. This h suggests that 50 expert days yields $z = 50$ at $\varepsilon = 0$ and $z = 250$ at $\varepsilon = 0.5$; investigation into the precise character of an appropriate h will be a necessary. For the specified h in our example we calculate the optimal operating point for the edge tracing algorithm, obtaining $\varepsilon^* \approx 0.247$ and resulting in $h(\varepsilon^*) \approx 157$ and $\beta(\varepsilon^*) \approx 0.599$. Thus optimizing the quantity/quality trade-off has yielded an improvement in power of almost 40%. We should engineer our edge tracing to operate at error rate $\varepsilon^* \approx 0.247$.

A summary of this example is presented in Figure 1.

Discussion

We conclude that we can indeed do a meaningful prospective power analysis for this connectome inference task, and that analysis of the quantity/quality trade-off between error in edge tracing and the number of putative edges traced is imperative for optimal neuroscientific experimental design.

The significance of our “admittedly idealized” illustrative setting is a simple version of a general question of scientific interest: how does connectivity probability depend on the neurons in question? Real scientific interest lies in more elaborate graph models and hypotheses – $K > 2$ kinds of cells and K^2 connection probabilities, or even an unknown number of cell types. The method described here can be generalized to these more realistic settings – some maintaining analytic tractability, but many realistic complex generalizations will of course require us to resort to numerical approximation methods. In hypothesis testing, one wishes to collect data in such a way so as to maximize the probability of rejecting the null hypothesis given that it is false. Often, the data collector is limited to only experimental intuition in making the quantity/quality decision. In some cases, however, one can turn to statistical connectomics to shed light on the quantitative trade-offs one expects with regard to a particular statistical inference question. Specifically, we have demonstrated that one can approximate the optimal operating point for the (errorful) edge tracing algorithm.

The above example uses statistical connectomics to address an important decision in neuroscientific data collection and analysis. While the results presented apply to a special case, general lessons can be learned.

In particular, we see that explicitly modeling the quantity/quality trade-off can yield significant inferential advantages. Note that the optimal operating point depends heavily on specific model assumptions; thus, any conclusions from such a prospective analysis are subject to the adequacy of those assumptions. We emphasize that this analysis is fundamentally a function of the particular inference task. Although we have outlined analytical results for one specific (i) inference task, (ii) graph model, (iii) error model, and (iv) quantity/quality trade-off function, each of these components must be customized for the neuroscientific question at hand.

The example results presented herein depend on knowing the quantity/quality function $z = h(\varepsilon)$. In general, this function will not be known, but it can be estimated. Specifically, consider the scenario of manual annotation of EM data. The performance of trained edge tracers operating so as to target various putative edge counts can be calibrated against a “gold” standard – derived, perhaps, using independent, complementary imaging methods – providing an estimate of h . As the size of connectome data sets continues to increase, the number of manual annotators required to estimate massive connectomes gets impractically large. Therefore, we will rely on machine vision algorithms to annotate the data (cf. the Open Connectome Project). The quantity/quality trade-off applies to such algorithms as surely as it applies to manual annotation, and the quantity/quality function h will need to be estimated.

The implications of optimizing the quantity/quality trade-off in connectome inference are potentially substantial in light of the recent global investment in connectome science. For example, the USA National Institute of Health (NIH) has budgeted over \$30 million to the Human Connectome Project, which aims to collect and analyze human magnetic resonance (MR) connectomes. Similarly, the European Union (EU) is potentially granting up to €1 billion to the Human Brain Project. Understanding and exploiting the quantity/quality trade-off in connectome inference will be essential to the efficient use of the available resources.

References

- Davi D. Bock, Wei-Chung A. Lee, Aaron M. Kerlin, Mark L. Andermann, Greg Hood, Arthur W. Wetzel, Sergey Yurgenson, Edward R. Soucy, Hyon S. Kim, and R. Clay Reid. Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471(7337):177–182, March 2011.
- V Braitenberg and A Schüz. *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer, Berlin, Germany, 1998.
- Human Brain Project. URL <http://www.thehumanbrainproject.com/>.
- Human Connectome Project. URL <http://www.humanconnectomeproject.org/>.
- Open Connectome Project. URL <http://www.openconnectomeproject.org/>.

β and $\partial\beta$

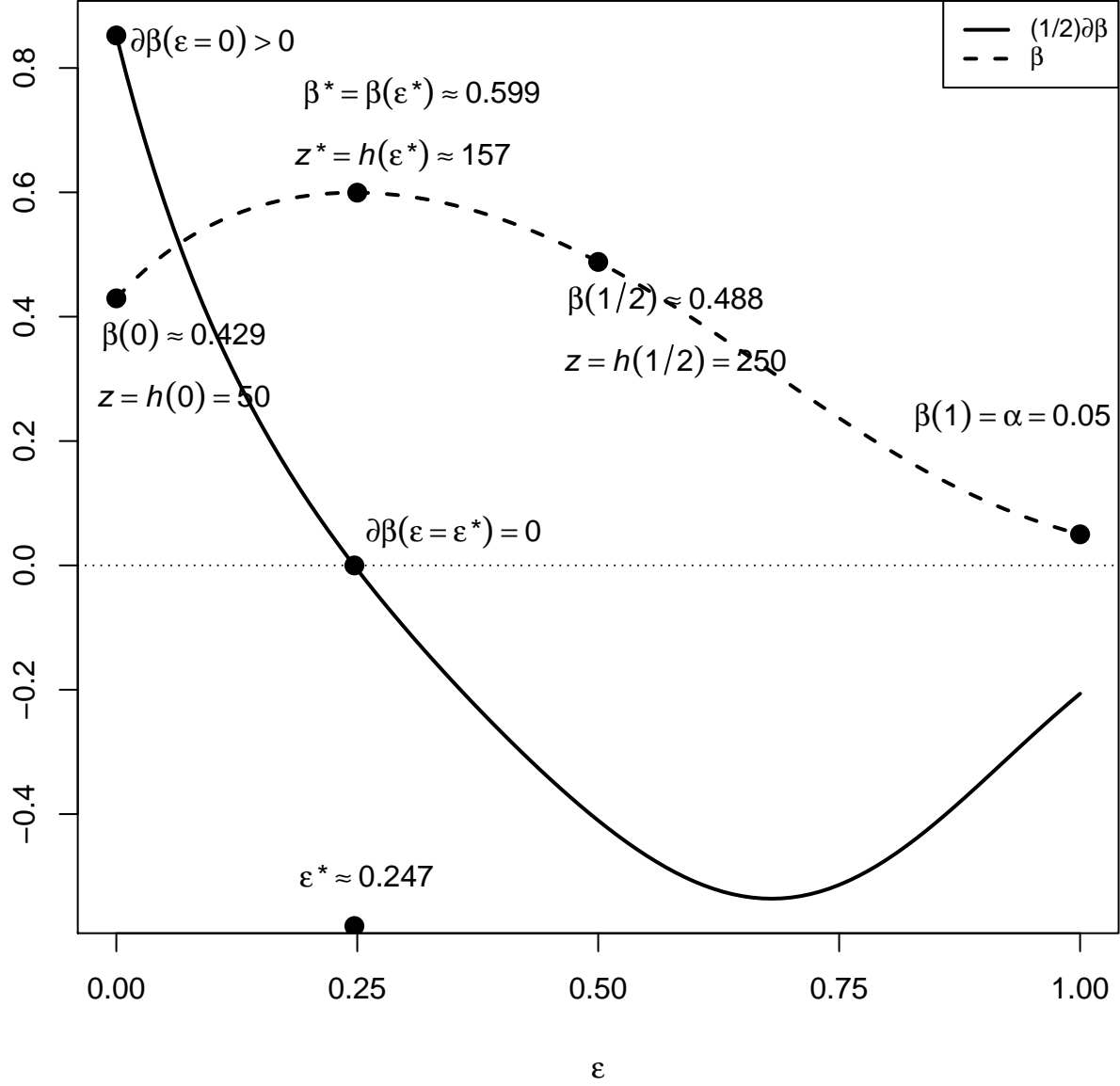


Figure 1: Power β and its derivative $\frac{\partial\beta}{\partial\epsilon}$ as functions of the edge tracing error rate ϵ for our example scenario (see text for details). (We plot $(\frac{1}{2})\frac{\partial\beta}{\partial\epsilon}(\epsilon)$ so that the two curves are on approximately the same scale and can productively be presented on the same plot.)